BARD Final Scientific Report
Cover Page

Date of Submission of the report: Dec 10 2005

BARD Project Number: IS-3454/03

Project Title:
Harnessing the genetic diversity engendered by alternative gene splicing

Investigators	Institutions
Principal Investigator (PI):	
Robert Fluhr,	*Weizmann Institute*
Co-Principal Investigator (Co-PI):	
Volker Brendel,	*Iowa State University*

Collaborating Investigators:

Keywords *not* appearing in the title and in order of importance. Avoid abbreviations.
Data base, Bioniformatics, Arabidopsis, rice, eudicots, monocots

Abbreviations commonly used in the report, in alphabetical order**:**
AS, alternative splicing

Budget: IS: $170.000 US: $130.000 Total: $300.000

Robert Fluhr

Signature
Principal Investigator

Signature
Authorizing Official, Principal Institution

BARD Final Scientific Report
Cover Page

Publication Summary (numbers)

	Joint IS/US authorship	US Authors only	Israeli Authors only	Total
Refereed (published, in press, accepted) BARD support acknowledged		3	2	5
Submitted, in review, in preparation			1	1
Invited review papers				
Book chapters				
Books				
Master theses			1	1
Ph.D. theses		2	1	3
Abstracts			2	2
Not refereed (WEB Site, proceedings, reports, etc.)		1		1

Postdoctoral Training: List the names and social security/identity numbers of all postdocs who received more than 50% of their funding by the grant.

Cooperation Summary (numbers)

	From US to Israel	From Israel to US	Together, elsewhere	Total
Short Visits & Meetings				0
Longer Visits (Sabbaticals)				0

Description of Cooperation

In this bioinformatics project, cooperation was achieved through the WEB site data transfer created specifically by this project. The Fluhr lab has made use of the sequence search methodology to identify potential exon/intron structure in pre-mRNA by splice site prediction and spliced alignment http://www.plantgdb.org/cgi-bin/PlantGDB/GeneSeqer/PlantGDBgs.cgi and the database of alternative splicing events in Arabidopsis available at AtGDB (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/). EST cluster information as well as Arabidopsis global unpublished EST alignments were shared between the labs.

Patent Summary (numbers)

	Israeli inventor only	US inventor only	Joint IS/US inventors	Total
Submitted				0
Issued (allowed)				
Licensed				0

BARD Final Scientific Report
Cover Page

Abstract

Our original objectives were to assess the unexplored dimension of alternative splicing as a source of genetic variation. In particular, we sought to initially establish an alternative splicing database for *Arabidopsis*, the only plant for which a near-complete genome has been assembled. Our goal was to then use the database, in part, to advance plant gene prediction programs that are currently a limiting factor in annotating genomic sequence data and thus will facilitate the exploitation of the ever increasing quantity of raw genomic data accumulating for plants. Additionally, the database was to be used to generate probes for establishing high-throughput alternative transcriptome analysis in the form of a splicing-specific oligonucleotide microarray.

We achieved the first goal and established a database and web site termed Alternative Splicing In Plants (ASIP, http://www.plantgdb.org/ASIP/). We also thoroughly reviewed the extent of alternative splicing in plants (Arabidopsis and rice) and proposed mechanisms for transcript processing. We noted that the repertoire of plant alternative splicing differs from that encountered in animals. For example, intron retention turned out to be the major type. This surprising development was proven by direct RNA isolation techniques. We further analyzed EST databases available from many plants and developed a process to assess their alternative splicing rate. Our results show that the lager genome-sized plant species have enhanced rates of alternative splicing. We did advance gene prediction accuracy in plants by incorporating scoring for non-canonical introns. Our data and programs are now being used in the continuing annotation of plant genomes of agronomic importance, including corn, soybean, and tomato.

Based on the gene annotation data developed in the early part of the project, it turned out that specific probes for different exons could not be scaled up to a large array because no uniform hybridization conditions could be found. Therefore, we modified our original objective to design and produce an oligonucleotide microarray for probing alternative splicing and realized that it may be reasonable to investigate the extent of alternative splicing using novel commercial whole genome arrays. This possibility was directly examined by establishing algorithms for the analysis of such arrays. The predictive value of the algorithms was then shown by isolation and verification of alternative splicing predictions from the published whole genome array databases.

The BARD-funded work provides a significant advance in understanding the extent and possible roles of alternative splicing in plants as well as a foundation for advances in computational gene prediction. This work has put the study of alternative splicing in plants on a firm basis both from a practical point of view (e.g. available interactive data bases), from a theoretical point of view (e.g. improving gene prediction programs and showing plant-specific splicing choice) and by extrapolating the knowledge obtained from model plants to agriculturally important species.

BARD Final Scientific Report
Cover Page

Summary of achievements

Our achievements were documented in three publications (see list below). We provide brief abstracts below. As a whole, our BARD-funded work provides a significant advance in understanding the extent and possible roles of alternative splicing in plants as well as a foundation for advances in computational gene prediction. These achievements will become particularly important as further crop genomes will become available for analysis in the coming years.

Publication 1

Motivation: The vast majority of introns in protein-coding genes of higher eukaryotes have a GT dinucleotide at their 5'-terminus and an AG dinucleotide at their 3'-end. About 1-2% of introns are non-canonical, with the most abundant subtype of introns being characterized by GC and AG dinucleotides at their 5'- and 3'-termini, respectively. Most current gene prediction software, whether based on *ab initio* or spliced alignment approaches, does not include explicit models for non-canonical introns or may exclude their prediction altogether. With present amounts of genome and transcript data it is possible to apply statistical methodology to non-canonical splice site prediction. We pursued one such approach for the training and implementation of GC-donor splice site models for *Arabidopsis* and rice. Our results indicated that the incorporation of non-canonical splice site models yields dramatic improvements in annotating genes containing GC-AG and AT-AC non-canonical introns. Comparison of models shows differences between monocot and dicot species, but also suggests GC-intron specific biases independent of taxonomic clade. We also present evidence that GC-AG introns occur preferentially in genes with atypically high exon counts.

Publication 2

U2AF is an essential splicing factor with critical roles in recognition of the 3'-splice site. In animals, the U2AF small subunit (U2AF^{35}) can bind to the 3'-AG intron border and promote U2 snRNP binding to the branchpoint sequences of introns through interaction with the U2AF large subunit. Two copies of U2AF^{35}-encoding genes were identified in *Arabidopsis* (atU2AF^{35}a and atU2AF^{35}b). Both are expressed in all tissues inspected, with atU2AF^{35}a expressed at a higher level than atU2AF^{35}b in most tissues. Differences in the expression patterns of atU2AF^{35}a and atU2AF^{35}b in roots were revealed by a promoter::GUS assay, with atU2AF^{35}b expressed strongly in whole young roots and root tips and atU2AF^{35}a limited to root vascular regions. Altered expression levels of atU2AF^{35}a or atU2AF^{35}b cause pleiotropic phenotypes (including flowering time, leaf morphology, and flower and silique shape). Novel slicing isoforms were generated from FCA pre-mRNA by splicing of non-canonical introns in plants with altered expression levels of atU2AF^{35}. U2AF^{35} homologs were also identified from maize, rice and other plants with large-scale EST projects. A novel C-terminal domain (SERE) is highly conserved in all seed plant protein homologs, suggesting it may have an important function specific to higher plants.

Publication 3

Alternative splicing (AS) has been extensively studied in mammalian systems, but much less in plants. In this paper we reported AS events deduced from EST/cDNA analysis in two model plants: Arabidopsis and rice. In Arabidopsis, 4,707 (21.8%) of the genes with EST/cDNA evidence show 8,264 AS events. About 56% of these events are intron retention (IntronR), and only 8% are exon skipping (ExonS). In rice, 6,568 (21.2%) of the expressed genes display 14,542 AS events, of which 53.5% are IntronR and 13.8% are ExonS. The consistent high frequency of IntronR suggests prevalence of splice site recognition by intron definition in plants. Different AS events within a given gene occur for the most part independently. 22%-30% of the AS events occur in untranslated regions (UTRs), and 14-16% of the AS events maintain the open reading frame (ORF) relative to the constitutive transcript. The remaining AS events change the start and/or stop codon position. In total, 36-43% of the AS events produce transcripts that would be targets of the nonsense-mediated decay (NMD) pathway, if that pathway were to operate in plants as in humans. 40% of Arabidopsis AS genes are also alternatively spliced in rice, with some examples strongly suggesting a role of the AS event as an evolutionary conserved mechanism of post-transcriptional regulation. A small number of previously un-annotated AS events in Arabidopsis were experimentally confirmed by RT-PCR. We created a comprehensive web-interfaced database to compile and visualize the evidence for alternative splicing in plants (ASIP, available at: http://www.plantgdb.org/ASIP/).

Publication 4

Alternative splicing combines different transcript splice junctions that result in transcripts with shuffled exons, alternative 5' or 3' splicing sites, retained introns and different transcript termini. In this way, multiple mRNA species and proteins can be created from a single gene expanding the potential informational content of eukaryotic genomes. Search algorithms of alternative splicing forms in a variety of *Arabidopsis* databases showed they contained an unusually high fraction of retained introns (above 30%), compared to 5% that was reported for humans. The preponderance of retained introns (65%) were either part of open reading frames, present in the UTR region or present as the last intron in the transcript, indicating that their occurrence would not participate in nonsense-mediated decay. Interestingly, the functional distribution of the transcripts with retained introns is skewed towards stress and external/internal stimuli-related functions. A sampling of the alternative transcripts with retained introns were confirmed by RT-PCR and were shown to co-purify with polyribosomes, indicating their nuclear export. Thus, retained introns are a prominent feature of alternative splicing in *Arabidopsis* and as such may play a regulatory function.

Publication 5

Alternative splicing (AS) is an important post-transcriptional regulatory mechanism that can increase protein diversity and affect mRNA stability. Different types of alternative splicing have been observed; these include exon skipping, alternative donor or acceptor site, and intron retention. In humans, exon skipping is the most common type while intron retention is the rare. In contrast, in *Arabidopsis*, intron retention is the most prevalent AS type (~40%). Here we show that direct transcript expression analysis using high-density oligonucleotide-based whole-genome microarrays (WGAs) is particularly amenable for assessing global intron retention in *Arabidopsis*. By applying a novel algorithm retained introns are detected in 8% of the transcripts examined. A sampling of 14 transcripts showed that 86% can be confirmed by RT-PCR. This rate of detection predicts an overall total AS rate of 20% for *Arabidopsis* compared to 10-22% based on EST/cDNA-based analysis. These findings will facilitate monitoring constitutive and dynamic whole genome splicing on the next generation WGA slides.

Publication 6

Alternative splicing (AS) can add significantly to genome complexity. Plants are thought to exhibit less alternative splicing then animals. An algorithm, based on EST pairs gapped alignment (EPGA), was developed that takes advantage of the relatively small intron and exon size in plants and directly compares pairs of ESTs to search for alternative splicing. EPGA was tested in *Arabidopsis* and rice for which annotated genome sequence is available and was shown to accurately predict splicing events. The method was applied to 14 plant species that include 24 cultivars for which enough ESTs are available. The results show up to a 4-fold difference in AS rates between plant species, with Arabidopsis and rice in the lower range and lettuce and sorghum in the upper range. Hence, compared to higher animals, plants show a much greater degree of variety in their AS rates and in some plant species the rates of animal and plant AS are comparable. In eudicots but not monocots, a correlation between genome size and AS rates was detected implying that the mechanisms that lead to larger genomes are a driving force for the evolution of AS.

Web site (http://www.plantgdb.org/ASIP/)

Cooperation was achieved through utilization of processed data. The Fluhr lab has made use of the sequence search methodology to identify potential exon/intron structure in pre-mRNA by splice site prediction and spliced alignment http://www.plantgdb.org/cgi-bin/PlantGDB/GeneSeqer/PlantGDBgs.cgi and the database of alternative splicing events in Arabidopsis available at AtGDB (http://www.plantgdb.org/AtGDB/prj/ZSB03PP/). EST cluster information as well as Arabidopsis global unpublished EST alignments were shared between the labs.

BARD Final Scientific Report
Cover Page

Appendix

Publications crediting the BARD grant

1. Sparks, M.E. & Brendel, V. (2005) Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics* **21 Suppl. 3**, iii20-iii30.
2. Wang, B.-B. & Brendel, V. (2006) Molecular characterization and phylogeny of U2AF1 homologs in plants. *Plant Physiol.* **140**, 624-636.
3. Wang, B.-B. & Brendel, V. (2006) Genome-wide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA* **103**, 7175-7180.
4. Ner-Gaon, H, Halachmi, R, Savaldi-Goldstein, S, Rubin, E, Ophir, R & Fluhr R (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.* **39**, 877-885.
5. Ner-Gaon, H & Fluhr R (2006) Whole Genome microarray in *Arabidopsis* facilitates global analysis of retained introns. DNA Res. 13: 111-121.
6. Ner-Gaon H, Leviatan N, Rubin E, & Robert Fluhr (2006) Comparative cross-species alternative splicing in plants (submitted).
7. Web site: http://www.plantgdb.org/ASIP/